DOCUMENT RESUME

ED 086 575 SO 006 685

AUTHOR	Smith, Leon I.; Greenberg, Sandra
TITLE	Test Use and Test Reliability in a Curriculum for Educable Mentally Retarded Children. Working Paper Number 1.
INSTITUTION	Yeshiva Univ., New York, N.Y. Curriculum Research and Development Center in Mental Retardation.
SPONS AGENCY	Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C.
BUREAU NO	6-1368
PUB DATE	Apr 73
NOTE	28p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Curriculum Evaluation; *Educable Mentally Handicapped; Educational Research; Evaluation Methods; *Evaluation Techniques; Measurement Techniques; Primary Grades; Reliability; Research Methodology; Skill Development; Statistical Analysis; Testing; *Testing Problems; *Test Reliability
IDENTIFIERS	*Social Learning Curriculum

ABSTRACT

A discussion of selected applications of new tests developed within the context of a large-scale curriculum for educable mentally retarded (EMR) children, the Social Learning Curriculum (SLC), is presented in this paper which investigates three types of reliability that need to be demonstrated in order to provide a basis of these applications. The three reliability coefficients refer to differences among students, classrooms, and tests. The SLC model is based on social environment levels (the Self, the Home and Family, the Neighborhood, and the Community). The SLC Survey Test is an experimental set of test items developed in an effort to tap samples of the conceptual skills contained in each of the 11 phases of the Self level. For this study, five phase tests were randomly selected and administered to ten randomly selected students from two samples of EMR primary classrooms and, in the analysis of data provided, three reliability coefficients derived. A discussion of the over-all findings suggests that the test measures are not adequate for their purposes at their present level of development. References and tables are included. Related documents are ED 075 972, ED 084 658, SO 006 684, 686 and 688. (Author/KSM)

ED 086575

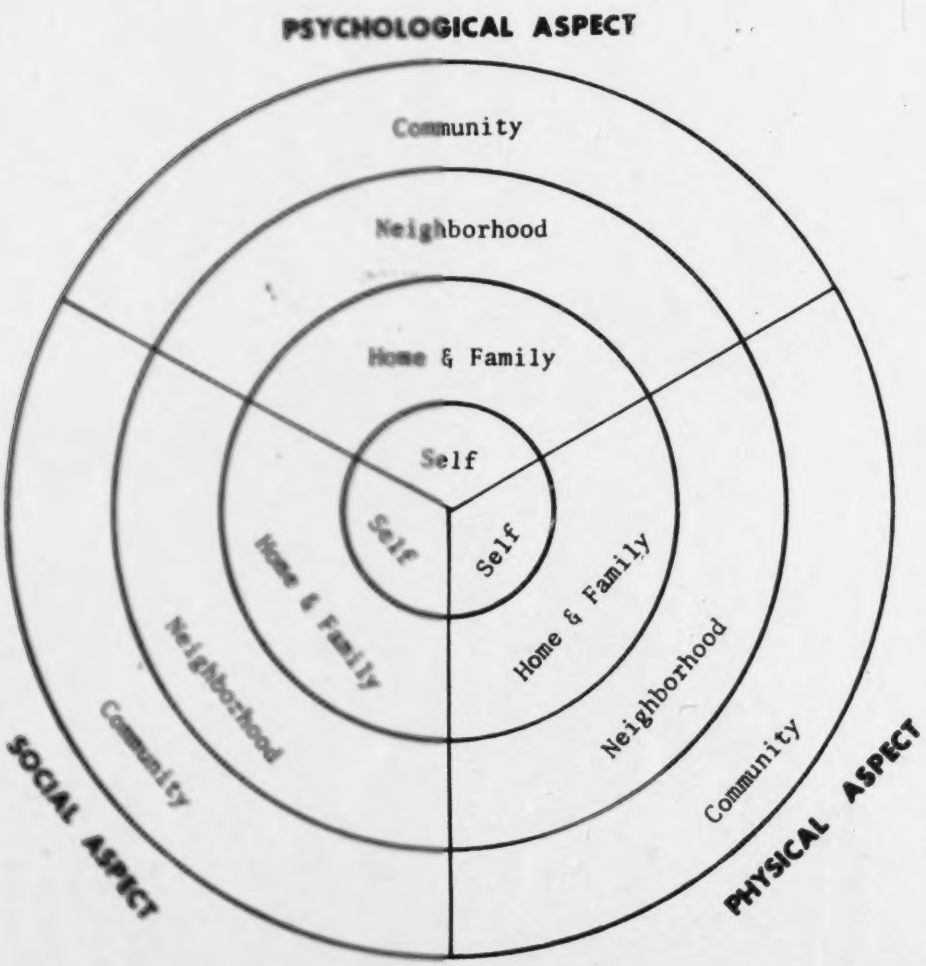
SP006 685

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Herb Goldstein
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER

A Working Paper

Test Use and Test Reliability in a Curriculum For Educable Mentally Retarded Children[1]

I. Leon Smith and Sandra Greenberg



Social Learning Curriculum
F.G.S. Yeshiva University, New York, N.Y.
Herbert Goldstein, Director

April 1973

FILMED FROM BEST AVAILABLE COPY

Curriculum Research and Development Center Staff

Director	Dr. Herbert Goldstein
Development Staff	Sarah Oelberg, Development Program Coordinator Dr. Philip Reiss, Development Program Coordinator Karen Kass, Development Program Specialist Patricia Charmoy, Development Program Specialist Mark Alter, Development Project Specialist Jerry Bernstein, Development Project Specialist Nancy Chaikin, Development Project Specialist
Field Operations Staff	Marjorie Goldstein, Field Operations Coordinator Michael Berlin, In-Service Training Program Specialist
Media Staff	David Bocher, Media Coordinator Jeffrey Telles, Media Specialist
Research Staff	Dr. I. Leon Smith, Research Coordinator Gregory Schimoler, Computer Specialist Steven Singer, Research Specialist
Evaluation Staff	Dr. Barry Lehrer, Evaluation Coordinator Sandra Greenberg, Evaluation Specialist Joyce Warshow, Evaluation Specialist Raymond Bapko, Evaluation Specialist
Administrative Staff	Dale Zevin, Administrative Secretary Rose Friedman, Secretary Alice Hall, Secretary Harry Kavee, Secretary Rochelle Mohr, Secretary
Editor	Caroline Cordts

Curriculum Research and Development Center in Mental Retardation
Department of Special Education
Ferkauf Graduate School of Humanities and Social Sciences
Yeshiva University
55 Fifth Avenue
New York, New York 10003

The Curriculum Research and Development Center is supported by the U.S. Office of Education, Bureau for the Education of the Handicapped (HEW) Project No. 422309

(c) Copyright - 1969 - Curriculum Research and Development Center in Mental Retardation, Yeshiva University, New York

Test Use and Test Reliability in a Curriculum For Educable Mentally Retarded Children[1]

I. Leon Smith[2] and Sandra Greenberg
Curriculum Research and Development
Center in Mental Retardation
Yeshiva University

Working Paper No. 1

Evaluation research is one of the deities invoked by educators to determine the utility of innovative approaches through the collection of hard data about their performance. While the process is universally praised by curriculum developers and researchers alike, the good works done in its name are remarkably few. Evaluation, in fact, is widely regarded as the least satisfactory component of program development. Guba (1969), for example, refers to the decades of non-significant differences that have been produced by the application of evaluation procedures to comparative studies of alternatives in all fields of education. Unfortunately, it is impossible to tell whether the absence of significant differences is a result of the failure of the evaluation procedures and the measures employed to detect differences or the inability of the programs to produce the desired effects, or both. Much of the difficulty could be eliminated by initially examining the ability of the instruments to detect specific kinds of differences before they are put to use for evaluation and research purposes. Tests that do not detect certain differences should not be used in curriculum applications where those differences are being investigated. This issue is particularly crucial in curriculum evaluation and research where new instruments must be developed because available, standardized measures are not substantively and methodologically appropriate.

The purpose of this paper is to discuss selected applications of new tests developed within the context of a large-scale curriculum for educable mentally retarded (EMR) children, namely, the Social Learning Curriculum (Goldstein, 1969; Heiss & Mischio, 1971), and to investigate three types of reliability that need to be demonstrated in order to provide a basis for these applications. The three reliability coefficients refer to differences among students, classrooms, and tests.

Applications Based on Student Differences

One anticipated use of the tests is the more homogeneous regrouping of EMR children within existing special classes for the purpose of providing more adequate instruction based on the Social Learning Curriculum (SLC). This approach differs considerable from traditional ability grouping that forms the basis for much of the recent discussion concerning the adequacy of the special class concept (MacMillan, 1971; Dunn, 1968). Under the SLC-test grouping approach, EMR children would be placed together based on behaviors that are specifically related to the content of instruction rather than on the basis of IQ alone.[3]

Although a number of potentially useful grouping algorithms are available for this purpose (Johnson, 1967; Cole, 1969; McQuitty, 1960, 1970), the statistical properties of the procedures are such that groupings of students are generated regardless of the quality (reliability) of the measuring instruments (Baker, 1972). Thus, before grouping procedures are applied, the reliability of the SLC measures with respect to differentiating students must be demonstrated. The use of high quality data in itself, however, provides no guarantee that the results will be more than nonsense or that the generated

groupings can be translated into different instructional methodologies and teacher behaviors. Recent evidence on this point from outside the field of special education suggests that the probability of obtaining meaningful groups is increased when task-specific achievements or measures of behavior directly related to the outcomes of instruction are employed as opposed to more general measures (Gagné & Gropper, 1965). Within the field of special education, Clausen (1972) reported that an attempt to define sub-groups of mental deficiency on the basis of constellations of basic abilities in sensory, motor, perceptual, and cognitive functions was not particularly successful. These findings, then, are consistent with the use of SLC-based tests for the purpose of regrouping EMR children.

Other attempts to form groups that extend beyond the use of IQ are discussed by MacMillan & Jones (1972), Jordan (1971), and Leland (1972). In these approaches, however, the variables or behaviors employed to group students are not related to or derived from a specific curriculum. Furthermore, given the importance of pupil grouping to present practices within the field of special education, it is surprising that these approaches tend to rely on judgmental and impressionistic combinations of test scores and pupil characteristics when well developed grouping algorithms of the type referred to in this paper exist.

Another test application for which reliable student defferences need to be demonstrated concerns the exploration of relationships with other individual difference variables. The latter are likely to include variables considered to be more direct measures of the characteristics in question such as observed behaviors in classroom settings (criterion-related validity) in addition to other constructs such as IQ and measures of perceptual motor ability (construct validity). In this connection, an important validational consideration involves

the degree to which the methodology employed in the development of the SLC tests accomplishes its purpose, namely, to measure knowledge of certain social concepts while minimizing hypothesized deficits and difficulties associated with the assessment of retardates' performance. A more complete discussion of this issue in relation to the SLC tests is presented in the method section of the paper. This line of inquiry should lead to the notion of aptitude-assessment interactions (AAIs) as the testing analogue to the investigation of aptitude-treatment interactions (ATIs). The ATI position states that, given a common set of instructional objectives, some students will be more successful with one type of instruction, while other students will be more successful with an alternate program (Bracht, 1970). The AAI view suggests that, given the same set of instructional objectives, some students will demonstrate the behaviors more successfully on one type of test, while other students will be more successful in performance on an alternative type of test.

Two heuristic testing models appear useful for the generation of AAIs within the context of the SLC.[4] One model might attempt to compensate in the testing situation for learner deficits or deficiencies presumed to be related to test performance by providing thos conditions that the EMR child cannot supply for himself. The actual deficit or deficiency is left untouched, and only the debilitating effects are circumvented through the design of the test and/or the testing situation. As an alternate to the compensatory model, testing procedures can be developed to capitalize on the retardate's relative ability strengths. This type of model is isomorphic in the sense that the testing procedure is matched to one of the retardate's higher aptitudes or to an ability where there is no presumed deficit. Here again, no attempt is made to modify deficits through testing.

At their present level of development, the SLC tests are compensatory with respect to the motivational difficulties of retardates and isomorphic in relation to their verbal deficits. That is, the testing situation is designed to heighten motivation, while the test itself is pictorial in nature and assumes that the retardate's visual abilities are better than his verbal abilities. Since this assumption is not likely to hold for all retardates, it is anticipated that alternate testing procedures will be developed. Hopefully, this kind of work will produce results that can serve as the basis for the development of a variety of instructional strategies that parallel the assessment procedures. At a more general level, the investigation of AAIs within this framework should also have implications concerning the distinction between competence and performance that has been advanced in other contests (Cole & Bruner, 1971; Bortner & Birch, 1970).

Applications Based on Classroom Differences

One obvious use of the tests is to employ them as criterion variables in a complex multi-treatment or simple experimental-contrast group design in an effort to assess the impact of the SLC on student learning. Here, the reliability of student differences is not of concern because the appropriate unit of analysis is not the student but the classroom (Glass, 1967; Glass & Stanley, 1970; Wardrop, 1969; Raths, 1969; Page, 1965; Wiley, 1965). The reason for this rests with the type of instruction provided by the SLC, or for that matter, any program which is not completely individualized.[5] Since the SLC involves programming for all students in a class simultaneously, the responses of the students within a class are not independent. Furthermore, the lack of independence would occur whether individual students or intact classes

are randomly assigned to the different treatments (Peckham, et al., 1969; Glass, 1967). However, since intact classes can be expected to respond independently of each other, a valid analysis can be performed on the classroom means employing the classroom as the unit of analysis. For this application, then, the reliability of the tests with respect to differentiating classrooms must be examined. See Wiley (1970) for an extended discussion and critique of this position.

Knowledge of reliable differences between classrooms on SLC tests would also suggest potential variability in teacher behaviors as a function of these differences as well as encourage attention to the effects of possible differences in other moments or characteristics of classroom distributions, particularly the variance. For example, it might be suspected that classes with low achievement variances may be taught by teachers who spend a disproportionate amount of time with the more retarded students, while classes with higher achievement variances may be taught by teachers who focus on the able students (Peckham, et al., 1969). It is also possible that classrooms with low achievement and behavior variances may be characterized by a teacher-directed atmosphere, while classrooms with higher variances may have a student-directed atmosphere (Costin, 1971). Lohnes (1972) extends this point and defines a classroom's syntality as the behavioral characteristics of the individuals comprising it, and suggests that the syntality of the classroom distribution should include reference not only to the mean and variance on a measure, but its skewness and kurtosis as well. Thus, an examination of the relationship between the syntality of special classrooms employing the SLC and teacher behaviors might well provide a possible explanation of the results of other curriculum studies which reported wide

variations in the classroom behaviors of teachers using the same instructional materials (Gallagher, 1966; Rosenshine, 1970,1972). A more traditional, alternate working hypothesis might suggest that variations in the behaviors of teachers using SLC materials are not due to differences in the syntality of the classrooms but simply to differences in the teachers' characteristics, attitudes, and beliefs in relation to both their failure to initiate and maintain a high level of program implementation and the lack of an acknowledged body of pedagogy to which all teachers subscribe (Cohen, 1972).

Applications Based on Test Differences

For several test uses previously discussed, additional reliability information is needed concerning the degree to which differences can be detected among tests presumed to measure different SLC concepts. For example, knowledge of reliable differences among the tests would provide evidence of multidimensional structure and suggest that grouping procedures could be applied using each test separately with students being reconstituted based on the patterning of their scores. Failure to detect reliable differences among the tests would support a unidimensional view and lead to grouping procedures based on a total score on all of the SLC tests. This type of reliability also has implications for the design of curriculum evaluation studies; unidimensional results would indicate the use of a univariate design, while multidimensional findings would dictate the use of a multivariate design (Smith, 1972; Bock, 1966; Baker, 1969).

SUMMARY

This study investigated three types of reliability that need to be demonstrated in order to provide an empirical basis for the applications that have been discussed. Furthermore, by relating types of reliability to particular test uses, a clear guide is provided for the selection of appropriate reliability coefficients that is lacking in current treatments (APA, 1966; McGaw et al., 1972).

METHOD

The SLC and SLC-Based Tests

The pedagogical model of the SLC is based on the expansion of the growing individual's world through predominantly social environments or levels, namely, the Self, the Home and Family, the Neighborhood, and the Community. At the Self Level, facts about the child logically constitute the substance of learning. The teaching elements of the SLC at the Self Level are divided into 11 Phases, each dealing with an array of related concepts and associated behaviors. See Goldstein (1969) and Heiss & Mischio (1971) for an extended discussion of the rationale underlying the construction of the SLC.

The Social Learning Curriculum Survey Test (SLCST), an experimental set of test items, was developed in an effort to tap samples of the conceptual skills contained in each of the 11 Phases (Lehrer, Heiss & Mischio, 1971). The testing procedures reflect the need to assess retarded children in relation to specific objectives of the Self Level while minimizing their verbal deficits and motivational difficulties which have been reported to adversely affect performance (House & Zeaman, 1963; Garjuoy et al., 1967; Spreen, 1965; Luria, 1961; Green & Zigler, 1963; Stevenson & Zigler, 1957: Zigler, 1962).

The items require the student to listen to a question and respond by marking an "X" on one of four picture stimuli. The 11 Phase tests are prepared in separate booklets. In addition, there is one booklet containing practice items intended to provide the students with training in the format of the test. During testing, all instructions are read aloud by the test administrator. Instructions for each item are detailed and redundant in order to compensate for the poor verbal skills of the respondents. No reading skills are necessary to understand the instructions and, to avoid any possible confusion, no written instructions are included in the test booklets. For each item, the administrator holds up the test booklet so that the students can verify the page they should be working on, while a proctor circulates around the room encouraging the students to maintain their test-taking behavior.

For this study, the following five Phase tests of the SLCST were randomly selected:

1. Recognizing Dependence. These items are intended to measure the child's ability to identify various authority figures in the school setting and his understanding of their roles.

2. Recognizing and Reacting to Emotions. This test measures the child's ability to identify and differentiate various emotional states.

3. Communicating with Others. These items examine the child's ability to identify different modes of communication and to understand the symbols within various communication modalities. The test also assesses the ability of the child to relate appropriate communication modes to the feelings and moods of others, as well as his ability to choose the appropriate communication modality for different situational contexts.

4. Attaining Social Skills. These items are designed to measure the child's understanding of the appropriate behavioral responses to different social and

environmental situations.

5. Identifying Helpers. These items examine the ability of the child to know when to ask for help, whom to ask for help, how to ask for help, and what to do when problems arise in the classroom situation.

Two procedures were employed to reduce the original pool of SLCST items for each of the five selected Phases. First, items were eliminated that (a) did not, as originally intended, relate to the behavioral objectives of a particular Phase, (b) that contained vague, ambiguous picture stimuli, and (c) that possessed poor test characteristics (difficulty and discriminability) based on previous item analyses. Second, ten items per Phase were randomly selected from the remaining pool.

Samples

The subjects consisted of ten randomly selected students from each of 13 randomly selected primary level, EMR classrooms drawn from each of two geographical samples who have participated at various times during the last four years in the field testing of the SLC. See Fratkin (1972) for a complete description of field testing activities. The samples were selected to represent polarities in racial, ethnic, and social class composition. Sample A is located in predominantly white, working class communities in northeastern Pennsylvania where assignment to special class placement is the responsibility of one central agency. Sample B is located in southwestern Florida and represents a racially and economically heterogeneous population, including black and white English speaking families in addition to migrant, bilingual families. Assignment to special class is the responsibility of several placement agencies. The means and standard deviations for CA and IQ for both samples are presented in Table 1.

Insert Table 1 about here

Analysis of Data

A general data layout for the design is presented in Table 2..

Insert Table 2 about here

Estimates of the sources of variability needed to obtain the three reliabilities were generated based on a components of variance approach (Medley & Mitzel, 1963; Cronbach, et al., 1963; Cronbach, et al., in press; McGaw, et al., 1973; Lindquist, 1953). The procedure differs from traditional views of reliability in that it permits the simultaneous examination of many sources of variability employing an analysis of variance (ANOVA) model. An estimate of each component in the design was obtained from a completely random, partially nested three-way ANOVA with one subject per cell. The model for the analysis was

$$(1) \qquad X_{ijm} = \mu + C_i + S_{j(i)} + T_m + CT_{im} + ST_{j(i)m} + E_{ijm}$$

where μ is a general mean and E_{ijm} is specific error. The parentheses around the subscripts for the student (S) dimension indicates that it was nested within classes (C).

In the analysis, variability due to classes, students, and tests were considered systematic, while the other sources were allocated to error. This is consistent with the three types of reliability discussed in the first section of the paper. Thus, the model for the analysis can be re-written as

$$(2) \qquad X_{ijm} = \mu + C_i + S_{j(i)} + T_m + \varepsilon_{ijm}$$

In terms of partitioning the variance provided by the analysis then,

$$(3) \qquad \sigma^2_X = \sigma^2_C + \sigma^2_S + \sigma^2_T + \sigma^2_\varepsilon, \text{ where } (4) \quad \sigma^2_\varepsilon = \sigma^2_{CT} + \sigma^2_{ST} + \sigma^2_E.$$

Since there was only one observation per cell, $\sigma^2_{ST} + \sigma^2_E$ was estimated as $\hat{\sigma}^2_{RES}$. The expected mean squares were derived through procedures suggested by Cornfield & Tukey (1956) and are presented in Table 3, while the estimates of each component are contained in Table 4.

Insert Tables 3 and 4 about here

The analysis was performed by a computer program written by Finn (1971). Based on the results, three reliability coefficients were derived. The first provided a measure of the reliability with which special classes could be differentiated. This coefficient was estimated as

(5) $\hat{\rho}_C = \hat{\sigma}^2_C / \hat{\sigma}^2_C + \hat{\sigma}^2_\varepsilon$, where $\hat{\sigma}^2_C$ was estimated from Table 4 and σ^2_ε by substituting in equation (4) the appropriate estimates from Table 4.[6] The second coefficient provided an index of the reliability with which students could be distinguished in performance. This coefficient was estimated as .

(6) $\hat{\rho}_S = \hat{\sigma}^2_S / \hat{\sigma}^2_S + \hat{\sigma}^2_\varepsilon$.

Finally, the third coefficient indicated the degree to which tests measuring different social concepts could themselves be differentiated and was estimated as

(7) $\hat{\rho}^2_T = \hat{\sigma}^2_T / \hat{\sigma}^2_T + \hat{\sigma}^2_\varepsilon$.

Results

Estimates of each of the variance components as well as the three reliabilities for each sample are presented in Table 5.

Insert Table 5 about here

The over-all findings suggest that at their present level of development, the test measures are not adequate for the purposes discussed. That is, with respect to classrooms, differences in the average levels of performance are simply not apparent at either sample. This does not minimize the possibility of demonstrating substantial differences in the variances of the classes-a separate issue which will not be considered here. Although student differences appear to be the largest source of systematic variation at both samples, the reliability of these differences does not reach acceptable levels. Regarding the tests at the Self Level, the evidence suggests that a unidimensional view of the total score on the 5 phases is much more tenable than a multidimensional view. Finally, the data on the variance components from the two separately sampled areas appear quite comparable despite known differences in the racial, ethnic, and social class compositions of the samples.

Discussion

The failure to identify adequate between-class, -student, and -test differences may in part be due to the rationale underlying the development of the measures at the Self Level of the SLC. The major purpose was to construct items that tap objectives of the curriculum as opposed to items that simply discriminate. While heightened discriminability can certainly be achieved by deleting certain items and adding others, the procedure reduces the relevance of the items to specific curriculum objectives. Thus, the psychometric

criterion of test discriminability appears to be incompatable with the construction of items that are intended to assess content objectives (Husek,1969; Tyler,1966). At the same time, there does not seem to be much value in constructing curriculum-based measures for the purposes previously discussed that do not adequately distinguish the two units of analysis (students and classes). For example, the issue of the differential grouping of EMR children for the purpose of SLC instruction seems to require the very kind of discrimination that the data do not support, and perhaps, can never support based on the above procedures for developing and selecting items. It would appear, then, that different item selection procedures are needed depending on the test application. Two approaches warrant consideration. First, for those uses excluding the evaluation of the effectiveness of SLC instruction, the most reasonable model is one in which the items are sensitive to individual differences in the units of analysis. Wide latitude should be permitted in the development and selection of the items to insure adequate discriminability. In addition, the notion of social learning should be considered as a generic construct that is not limited to but extends beyond the specific objectives of the SLC. This approach requires the use of traditional, normative referenced test criteria for which the statistical analysis conducted in the present study is most appropriate.

Second, for the purpose of evaluating the outcomes of SLC instruction, the items should be sensitive to differences in instructional emphasis (Gleser, 1963; Hammock, 1960; Roudabush, 1973). For this use, item selection should conform as closely as possible to the specific objectives of the curriculum. The criteria employed here require a minimum two-stage tryout of items, that is, a pre-instruction administration of the items, followed by SLC instruction, then a

post-instruction administration of the items to the same students. For this use, it would be desirable to retain items which were responded to correctly by all those following instruction but were answered incorrectly prior to instruction. Since each approach follows different item selection procedures and different types of analysis, uniquely different tests are likely to be constructed. Recent evidence suggests that less than half the items selected by the normative criteria were selected based on the instructional criteria (Roudabush, 1973). The moral of this study appears to rest with the fact that item selection based primarily on the instructional model will not necessarily meet normative test criteria.

Superimposed on these considerations is the issue of the unit of analysis. If the present data are any relative indication, it will be far easier to differentiate EMR children than EMR classes at the same site under both item selection procedures. Furthermore, the items that discriminate individuals may not be the same ones that discriminate groups or classes (Lewy, 1973).

Footnotes

[1] The preparation of this paper was supported by a grant from the U. S. Office of Education, Bureau for the Education of the Handicapped, Project #6-1368.

[2] The authors wish to thank Herbert Goldstein, Barry Lehrer, and Gregory Schimoler for their helpful comments, and Carol Sternberg for her assistance in the data analysis.

[3] The idea of curriculum-based or instructional grouping discussed here could also be extended to include Thelen's (1967) notion of teachability grouping. This approach would require teachers to identify those students who did well in the class as well as those students who did not. The students would then be tested on variables relevant to effective classroom behavior, including the SLC-based tests. The responses that differentiate the two groups identified by the teacher could be made into a scoring key to be used with next year's classes. The teacher's "teachable" class, then, would be composed of high scorers using the compatability discriminating key.

[4] See Salomon (1972) for an extended discussion of the ATI analogues.

[5] For a discussion of instructional issues related to the differences between the student and the classroom as the unit of analysis see Thelen (1969) and Lindvall & Cox (1969).

[6] The estimate of classroom variation ($\hat{\sigma}^2_C$) also contained variability, if any,

due to differences among teachers and schools. Separate estimates of the three components can only be obtained in a design that samples at least two classrooms for each of at least two teachers at each of at least two schools. The design was impossible to implement in the present study since each teacher spends the entire day with the same self-contained special class, that typically being the only EMR class in the school.

References

American Psychological Association. Standards for educational and psychological tests and manuals. Washington, D. C.: APA, 1966.

Baker, R. L. Curriculum Evaluation. Review of educational research, 1969, 39 (3), 339-358.

Baker, F. B. Numerical Taxonomy for Educational Researchers. Review of educational research, 1972, 42 (3), 345-358

Bock, R. D. Contributions of Multivariate Experimental Designs to Educational Research. In R. B. Cattell (Ed.) Handbook of multivariate experimental psychology, Chicago: Rand McNally, 1966, pp. 820-840.

Bortner, M. & Birch, H. G. Cognitive Capacity and Cognitive Competence. American journal of mental deficiency, 1970, 74 (6), 735-744.

Bracht, G. H. Experimental Factors Related to Aptitude-Treatment Interactions. Review of educational research, 1970 40 (5), 627-645.

Clausen, J. The Continuing Problem of Defining Mental Deficiency. The journal of special education, 1972,6 (1), 97-106.

Cohen, E. G. Sociology and the Classroom: Setting the Conditions for Teacher-Student Interaction. Review of educational research, 1972, 42 (4), 441-452.

Cole, A. J. Numerical taxonomy. New York: Academic Press, 1969.

Cole, M. & Bruner, J. S. Cultural Differences and Inferences about Psychological Processes. American psychologist, 1971, 36 (10), 867-875.

Cornfield, J. & Tukey, J. W. Average Values of Mean Squares in Factorials. Annals of mathematical statistics, 1956, 27, 907-949.

Costin, F. Empirical Test of "Teacher Centered" versus "Student Centered" Dichotomy. Journal of educational psychology, 1971,62 (5), 410-412.

Cronbach, L. J., Rajaratnam, N. & Gleser, G. Theory of Generalizability; a Liberalization of Reliability Theory. British journal of statistical psychology, 1963, 16, 137-163.

Cronbach, L. J., Gleser, G. & Rajaratnam, N. Dependability of behavioral measurements. New York: Wiley, in press.

Dunn, L. M. Special Education for the Mildly Retarded-Is Much of it Justified? Exceptional children, 1968, 35 (1), 5-22.

Finn, J. D. Multivariance: univariate and multivariate analysis of variance, covariance, and regression. Ann Arbor, Michigan: National Educational Resources, 1972.

Fratkin, M. The Social Learning Curriculum. How to Use, Evaluate, and Field Test. Unpublished manuscript, Yeshiva University, 1972.

Gagné, R. M. & Gropper, G. L. Individual differences in learning from visual and verbal presentations. American Institutes for Research: Studies in Filmed Instruction, 1965.

Gallagher, J. J. Teacher Variation in Concept Presentation in BSCS Curriculum Program. Urbana: Institute for Research on Exceptional Children, Univ. of Illinois, 1966.

GerJuoy, I. R., Winters, J. J., Jr., Alvarez, J. M. & Pullen, M. M. Response Preference and Choice-Sequence Preference: II Perceptual and Motor Conditions. Psychonomic science, 1967, 8, 75-76.

Glass, G. V. The Experimental Unit and the Unit of Statistical Analysis: Comparative Experiments with Intact Groups. Institute for State Educational Agency Personnel, Denver, Col., 1967.

Glass, G. V. & Stanley, J. C. Statistical methods in education and psychology. Prentice-Hall, 1970.

Gleser, R. Instructional Technology and the Measurement of Learning Outcomes: Some Questions. American psychologist, 1963, 18, 519-521.

Goldstein, H. Construction of a Social Learning Curriculum. Focus on exceptional children, 1969, 1 (2), 1-10.

Green, C. & Zigler, E. Social Deprivation and the Performance of Retarded and Normal Children on a Satiation Type Task. Child development, 1962, 33, 499-508.

Guba, E. G. The Failure of Educational Evaluation. Educational technology, 1969, 9 (5), 29-38.

Hammock, J. Criterion Measures: Instruction vs. Selection Research. American psychologist, 1960, 15, 435.

Heiss, W. E. & Mischio, G. S. Designing Curriculum for the Educable Mentally Retarded. Focus on exceptional children, 1971, 3 (2), 1-10.

House, B. J. & Zeaman, D. Learning Sets from Minimum Stimuli in Retardates. Journal of comparative and physiological psychology, 1963, 56, 735-739.

Husek, T. T. Different Kinds of Evaluation and Their Implications for Test Development. Evaluation comment, 1969, 2, 8-10.

Johnson, S. C. Hierarchical Clustering Schemes. Psychometrica, 1967, 32, 241-254.

References (continued)

Jordan, J. B. Dial G for Grapevine: a Conversation in Exceptional Child Research. In J. B. Jordan & P. L. McDonald (Eds.) Dimensions: annual survey of exceptional child research activities and issues-1970. Arlington, Va.: The Council for Exceptional Children, 1971, 5-15.

Lehrer, B., Mischio, G. S. & Heiss, W. E. Manual: Social Learning Curriculum Survey Test for the Self Level. Unpublished manuscript, Yeshiva University, 1971.

Leland, H. Mental Retardation and Adaptive Behavior. The journal of special education, 1972, 6 (1), 71-79.

Lewy, A. Discrimination Among Individuals versus Discrimination Among Groups. Journal of educational measurement, 1973, 10 (1), 19-24.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953.

Lindvall, C. M. & Cox, R. C. Role of Evaluation in Programs for Individual Instruction. In R. W. Tyler (Ed.) Educational evaluation: new roles, new means, 1968 NSSE Yearbook, Chicago: Univ. of Chicago Press, 1969, 156-188.

Lohnes, P. R. Statistical Descriptors of School Classes. American educational research journal, 1972, 9 (4), 547-556.

Luria, A. F. The role of speech in the regulation of normal and abnormal behavior. New York: Liveright, 1961.

MacMillan, D. L. Special Education for the Mildly Retarded: Servant or Savant. Focus on exceptional children. 1971, 2 (9), 1-11.

MacMillan, D. L. & Jones, R. L. Lions in Search of More Christians. The journal of special education, 1972, 6 (1), 81-91.

McGaw, B., Wardrop, J. L. & Bunda, M. A. Classroom Observation Schemes: Where Are the Errors? American educational research journal, 1972, 9 (1), 13-27.

McQuitty, L. L. Hierarchical Syndrome Analysis. Educational and psychological measurement, 1960, 20, 293-304.

MCQuitty, L. L. Hierarchical Classifications by Multiple Linkage, Educational and psychological measurement, 1970, 30, 3-20.

Medley, D. M. & Mitzel, H. E. Measuring Classroom Behavior by Systematic Observation. in N. L. Gage (Ed.) Handbook of research on teaching. Chicago: Rand McNally, 1963, pp. 247-328.

Page, Ellis. B. Recapturing the Richness within the Classroom. Paper presented at the Annual meeting of the American Educational Research Association, Chicago, Ill., 1965.

Peckham, P. D., Gene V. Glass, Hopkins, K. D. The Experimental Unit in Statistical Analysis: Comparative Experiments with Intact Groups. Journal of special education, 1969, 3 (4), 337-349.

Raths, J. The Appropriate Experimental Unit, Educational leadership, 1967, 25, 263-266.

Rosenshine, B. Evaluation of Classroom Instruction, Review of educational research, 1970, 40 (2), 279-300.

Rosenshine, B. Translating Research into Action? Education leadership, 1972, 30, 594-597.

Salomon, G. Heuristic Models for the Generation of Aptitude-Treatment Interaction Hypotheses. Review of educational research, 1972, 42 (3), 327-343.

Smith, I. L. The ETA Coefficient in MANOVA. Multivariate behavioral research, 1972, 7, 361-372.

Spreen, I. Language Function in Mental Retardation: A Review: II Language in Higher Level Performance. American journal of mental deficiency, 1965, 70,351-362.

Stevenson, H. W. & Zigler, E. Discrimination Learning and Rigidity in Normal and Feebleminded Individuals, Journal of personality, 1957, 25, 699-711.

Thelen, H. A. Classroom grouping for teachability. New York: Wiley, 1967.

Thelen, H. A. The Evaluation of Group Instruction. In R. W. Tyler (Ed.) Educational evaluation: new roles, new means. 1968 NSSE Yearbook, part II, Chicago: Univ. of Chicago Press, 1969, 115-155.

Tyler, R. W. The Objectives and Plans for a National Assessment of Educational Progress. Journal of educational measurement, 1966, 1-10.

Wardrop, J. L. Controlled Experimentation in Multiclassroom Settings. Research and Development Strategies in Theory Refinement and Education Improvement. Theoretical Paper #15, Madison, Wis.; Research & Development Center for Cognitive Learning, Univ. of Wisconsin, 1968.

Wiley, D. E. Standard Experimental Designs and Experimentation with School Conditions. A paper presented at the annual meeting of the American Educational Research Association, 1965.

Wiley, D. E. & Bock, R. D. Quasi-experimentation in Educational Settings: Comment. The school review, 1968, 75, 353-366.

Wiley, D. E. Design and Analysis of Evaluation Studies. In M. C. Wittrock & D. C. Wiley (Eds.) The evaluation of instruction: issues and problems. New York: Holt, Rinehart, & Winston, 1970, 259-271.

Zigler, E. Social Deprivation in Familial and Organic Retardates. Psychological reports, 1962, 10, 370.

Table 1

Sample Characteristics

Measure	Sample A	Sample B
Chronological Age	$\overline{X}$ = 10.2	$\overline{X}$ = 9.5
	SD = 2.1	SD = 1.9
I Q [a]	$\overline{X}$ = 70.7	$\overline{X}$ = 66.4
	SD = 8.5	SD = 7.2

[a] estimates are based on the WISC or Stanford-Binet.

Table 2

Data Layout for Design

Factor C -Classes i = 1, 2, ... 13

Factor S (within C) -Students j = 1, 2, ... 10

Factor T -Tests m = 1, 2, ... 5

	C_1	..	C_i	..	C_{13}
	$S_{11} \cdot S_{j1} \cdot S_{10\ 1}$	..	$S_{1i} \cdot S_{ji} \cdot S_{10\ i}$	..	$S_{1\ 13} \cdot S_{j\ 13} \cdot S_{10\ 13}$
T_1					
:					
T_m			X_{ijm}		
:					
T_5					

Table 3

Expected Mean Squares

Source	E(MS)
Systematic	
Classes (C)	$\sigma^2 + \sigma^2_{ST} + 10\sigma^2_{CT} + 5\sigma^2_{S} + 50\sigma^2_{C}$
Students Within Class (S)	$\sigma^2 + \sigma^2_{ST} + 5\sigma^2_{S}$
Tests (T)	$\sigma^2 + \sigma^2_{ST} + 10\sigma^2_{CT} + 1300\sigma^2_{T}$
Error (σ^2_{ε})	
C x T	$\sigma^2 + \sigma^2_{ST} + 10\sigma^2_{CT}$
Residual	$\sigma^2 + \sigma^2_{ST}$

Table 4

Variance Components

Systematic		Components of Variance
$\hat{\sigma}^2_C$	=	$1/50\ (MS_C - MS_S - MS_{CT} + MS_{RES})$
$\hat{\sigma}^2_S$	=	$1/5\ (MS_S - MS_{RES})$
$\hat{\sigma}^2_T$	=	$1/130\ (MS_T - MS_{CT})$
Error ($\hat{\sigma}^2_\varepsilon$)		
$\hat{\sigma}^2_{CT}$	=	$1/10\ (MS_{CT} - MS_{RES})$
$\hat{\sigma}^2_{RES}$	=	MS_{RES}

Table 5

ANOVA for Components

Source	df	Sample A			Sample B		
		MS	$\hat{\sigma}^2$	$\hat{\rho}$	MS	$\hat{\sigma}^2$	$\hat{\rho}$
Classes (C)	12	11.5	.03	.02	17.3	.18	.10
Students (S)	117	7.3	1.2	.43	7.7	1.3	.44
Tests (T)	4	141.9	1.1	.41	66.1	.5	.20
C x T	48	4.3	.3		2.1	.06	
Residual	468	1.3	1.3		1.5	1.5	